Business Contracting Analysis

Keval Khara
kevalk@bu.edu
U05595501

Pranav Raikundalia
pranavpr@bu.edu
U67067969

Vaibhav Sharma
vaibhavs@bu.edu
U91099967

Background

We wouldn't have had the opportunity to work on this alarming situation if it wasn't for the New England Center for Investigative Reporting (NECIR) at Boston University, which seeks to expose injustice through investigative reporting and training. We also want to take this moment to thank Paul Singer, investigations editor at NECIR & WGBH, and Andrei Lapets, Associate Professor at Boston University, for their direction on this project.

NECIR stories reach wide audiences and spur action through mainstream news outlets such as the New York Times, Boston Globe and WGBH. NECIR reporters share their skills with student and mid-career journalists through workshops and classes to further expand NECIR's impact.

Problem Statement

To compare the data gathered by BU Spark from Massachusetts Central Register, with the state's databases of minority-owned and women-owned businesses, to look for patterns in the procurement database that may indicate the imbalance in participation by minority and women-owned businesses, as well as the disparity in the number of RFPs and the awards.

Datasets

We basically have 2 sources of data-

1. From BU-Spark! - http://necirspark.herokuapp.com

This data consists of all the contracts by the State and the businesses that they were awarded to. We can analyze the disparity between the number of RFPs and the awards from this database.

2. From https://www.sdo.osd.state.ma.us/BusinessDirectory/BusinessDirectoryDownload.aspx :
    a. SDO (MBE, WBE, PBE and NPO) Directory Listing by Business Name
    b. DBE Directory Listing by Business Name
    c. ACDBE Directory Listing by Business Name

    This data consists of information about all the minority and women-owned businesses in the state. We can compare this data with the contracts and their awards from the previous dataset to analyze the participation of minority and women-owned businesses.

In this report, we will primarily be working with the RFP and award datasets since we analyze the disparity in their numbers.

Approach

Our main challenge in this project was to deal with the dirty/empty and duplicate data in the given datasets. First, we basically produced cleaner datasets by removing all the duplicates and by not considering any RFPs where the project number was missing. We did this because we used the project numbers to match the RFPs and the awards from the two datasets.
We also calculated the number of awards each contractor was awarded. Something that we observed here was that there were many contractors with multiple similar names, which again happened because the data was not entered in a uniform manner. For example, *N.E.L Corporation, MA 01949* and *NEL Corporation, MA*, were considered as separate contractors. We needed to employ a separate algorithm to tackle this problem.

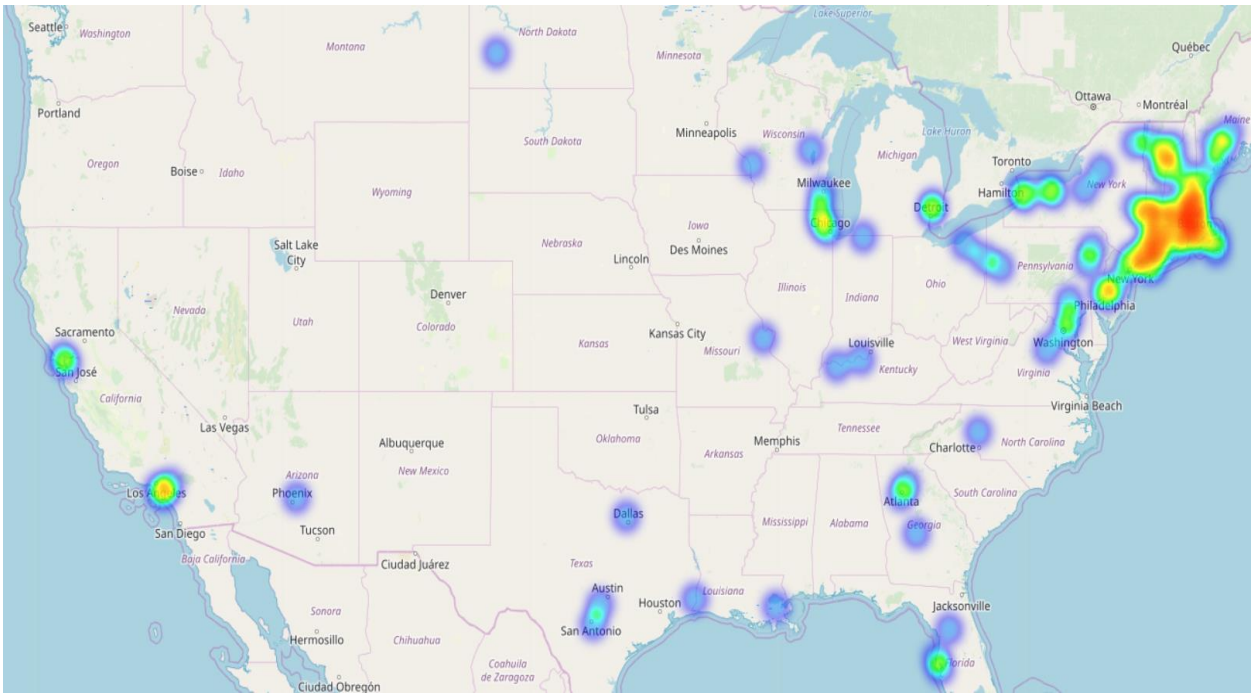
We also plotted a heatmap to see if there were any unusual clusters with respect to the contractors to whom the awards were given. We didn't observe any anomalies as such, apart from the fact that there is some concentration of these awards to contractors from the Boston and Rhode Island area, with a seemingly uniform distribution over all of Massachusetts. The contractor with the most awards is NEL Corporation (44), followed by Kurtz Inc. (41), both of which are based in Massachusetts.

Finally, we came across many conclusions which are mentioned in the next section. The key thing to keep in mind here is the fact that our conclusions might not be perfect in the sense that the data on which they are based might be skewed in the first place.

Conclusions

We produced the following datasets while tackling the problem-

1. **awards_per_contractor.csv -** All the contractors with their award count.
2. **intersection.csv** - Matched RFPs and awards
3. **unaccounted_RFPs** - All the RFPs that cannot be connected to an award
4. **heatmap.html** - Heatmap to visualize the distribution of awards to contractors



Heatmap highlighting the distribution of awards to contractors across the country

All the numbers-

- Total number of RFPs we found after removing the duplicates and empty/dirty data = 19,758

- Total number of awards we found after removing the duplicates and empty/dirty data = 7717

- Total number of RFPs we could match to awards = 5613

- Total number of RFPs that cannot be connected with any reasonable certainty to an award announcement = 14,145

- Total dollar amount of all the RFPs in our dataset = $41.91 Billion

- Total dollar amount of all the awards in our dataset = $10.73 Billion

- Total unaccounted “estimated cost” of RFPs (i.e. the estimated cost of all RFPs that we could not match with any award) = $23.26 Billion

- Total contract amount of all the awards we matched = $4.177 Billion

Future Work

We aim to provide real time analysis as and when the government database updates its content on awards and provide notifications on missing awards so that they can be dealt with immediately. This could be done with the help of a plugin or an API that integrates with the Massachusetts Central Register.

Furthermore, our findings are based on the data that is provided on the government website, which could be skewed in the first place. To validate all the data provided on the website would be a daunting but a useful task to get the best possible analysis.